1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 2(s) 호

한국어 문장 단위운율 발생에 관한 연구

A Study on the Prosody Generation of Korean Sentences

민경중* 이일구* 강찬구** 임운천*
*호서대학교 전자공학과, **안양과학대

Kyoung-Joong Min*, Il-Goo Lee, Chan-Goo Kang, Un-Cheon Lim
*Dept. of Electronics Eng., Hoseo University
**Anyang Science University

E-mail: uclim@dogsuri.hoseo.ac.kr

요약

법칙합성 시스템은 합성단위, 합성기, 합성방식 등에 따라 여러 가지 다양한 음성합성시스템이 있으나 순수한 법칙합성 시스템이 아니고 기본 합성단위를 연결하여 합성음을 발생시키는 연결합성 시스템은 연결단위사이 그리고 문장단위에서의 매끄러운 합성계수의 변화를 구현하지 못해 자연감이 떨어지는 실정이다. 자연감을 높이기 위해 보다 자연음에 가까운 운율을 발생시키기 위해 신경망으로 자연음의 운율을 발생시키기 위해 먼저 운율에 영향을 주는 요소들을 고려하여 신경망 입력 패턴을 구성한다. 분절요인에 의한 영향을 고려해주기 위해 전후 3음소를 동시에 입력시키고 문장내에서의 구문론적인 영향을 고려해주기 위해 해당 음소의 문장내에서의 위치, 운율구에 관한 정보등을 신경망의 입력 패턴으로 구성하였다. 신경망을 훈련시키기 위한 언어자료로는 고립단어군과 음소균형 문장군 그리고 삽입음절 연결어등으로 구성한다. 특정화자로 하여금 언어자료를 발성하게 한 음성시료의 운율을 분석하여 신경망을 훈련시켜 자연음의 운율과 유사한 합성운율을 발생시켰다.

Ⅰ. 서 론

음성합성은 인간의 음향학적 정보 전달수단인 음성을 기계가 소리의 합성을 통하여 발생시키는 기술이다. 이 기계에 의한 합성음은 올바른 정보 전달능력으로서 이해도와 인간의 발성과의 유사함을 나타내는 자연성으로 평가되어 진다. 또한 음성합성의 음성분야가 넓어지고 보편화됨에 따라, 인간의 음성과 같이 자연스러운 합성음에 대한 요구가 증가되고 있다.

법칙 합성 시스템은 합성단위, 합성기, 합성방식등에 따라 여러 가지 다양한 시스템이 있으나 순수한 법칙합성 시스템이 아니고 기본 합성단위를 연결하여 합성음을 발생시키는 연결합성 시스템은 연결단위사이에서의 매끄러운 합성계수의 변화를 구현하지 못해 자연감이 떨어지는 실정이다. 특히 시간영역 법칙합성 시스템의 합성음은 이해도는 향상되었음에도 불구하고 자연감이 많이 떨어지고 있다[8].

음성 언어자료는 상태에 따라 많은 데이터량을 가지고 있어, 이러한 데이터베이스를 신경망에 학습시키는데에는 많은 처리과정 및 시간이 필요하게 된다. 그리고 문장내에서 음절의 억양의 높고 낮음과 시간의 길고 짧은 경우도 각 음절마다 다 고려되어야 하며 접속되는 앞과 뒤의 음절에 따라 발음의 변화도 고려하여야 하며, 같은 음절이라도 음절위치에 따라서도 변화요인이 많기 때문에, 문장내에서 적절한 운율구의 경계를 추출하여, 이러한 경계의 전후 음절의 피치변화, 앞음절의 장음화, 음절변화등을 고려하고, 자연감을 위해서 평서문이 아닌 감탄문 또는 의문문과 같은 구내에서의 음절위치에 대한 변화 법칙을 고려한 언어자료을 구축하고 이를 토대로 음성시료를 채집하여 신경망을 학습시킴으로써 연속음에서 좀더 이해도 및 자연성을 향상시키는 방법을

제안하고자 한다.

Ⅱ. 운율제어

일반적으로 서로간 자연스러운 대화라든가 글을 읽을 때의 음성, 즉 자연음은 화자의 감정상태, 말의 내용 또는 강세, 발음속도와 같은 의미론적인 정보와 전체의 구문구조와 문장내에서의 구와 절의 경계위치, 기능, 상호 결합관계등의 구문론적인 정보, 단어에서의 강세유형과 각 분절음소의 전후 결합에 의한 영향이 특성을 결정하게 된다. 이와 같은 여러가지 요인에 의해 같은 음소일지라도 문장내에서의 위치에 의해 피치와 지속시간, 크기 등이 달라지는데 이들 피치와 지속시간 및 크기 등의 변화를 운율이라 한다.

피치의 변화에 영향을 주는 요인으로 먼저 의미론적인 요소 즉 대비, 강조, 화자의 감정상태와 발음속도등이 있고, 구문론적인 측면에서는 실제 대화체의 자연음에서의 피치변화를 모델링할 수 있는 단계에는 아직 미치지 못하고, 평서문에서 구문의 구,절 등의 경계와 단어의 강세유형 그리고 분절음소가 미치는 영향을 고려한 피치 모델을 구하고 있다.

지속시간을 변화시키는 요인으로 전후음소에 의한 영향, 강세의 유·무에 의한 영향, 휴지전 여부에 의한 영향, 음절수에 의한 영향, 단어의 빈도수에 의한 영향 등이 있다[2].

한국어의 지속시간에 대한 연구는 2음절로 구성된 무의미 단어에 대한 발음을 대상으로 하여 유성자음에 비해 무성자음 앞에 오는 단음의 지속시간이 짧아지고 마찰음에 비해 중지음 앞의 모음길이가 짧아짐을 보고하고 있고, 6음절 문장을 대상으로 하여 전체 문장 길이의 분산이 각 음절길이의 분산의 합에 비해 적어 음절이 기본단위가 아니고 분절이 기본단위임을 제시하고 있다.

에너지는 음소를 구별하는데 중요한 운율요소 중 하나로, 합성음의 자연성에 미치는 영향이 큰 강세 및 억양에 의해 변화하며 운율법칙의 중요한 요소이다[5].

의미론적인 정보에 의한 문장단위의 운율법칙을 추출하기 위해서는 여러 가지 상태에 따른 많은 양의 데이터와 오랜 처리시간이 필요하고 이러한 운율정보를 법칙화하는데 많은 어려움이 따르게 된다. 그러나 구문론적인 정보는 문장의 구조를 해석하여 구와 절 또는 운율구, 억양구 등의 경계를 분리해내어 구 단위의 운율을 생성하여, 문장단위에서 보다 쉽게 법칙을 추출하여 운율을 제어할 수가 있다.

Ⅲ. 문장내에서의 운율변화

Ⅲ-1. 강세

강세는 단어단위의 음성에서 앞뒤소리와 상대적인 관계에 의한 어느 한 음절을 강조함으로써 뜻의 전달에 기여하는데, 강세는 피치, 지속시간, 세기에 의해 구성되어진다. 일반적으로 강세는 한 문장의 표면구조인 한 개의 음형 강세로 되어 있다. 그리고 음향학적인면에서는 강세는 주위음절에 대한 특정음절의 F_0의 상승, 지속시간의 증가, 크기의 증가, 또는 이들의 복합현상으로 표현할 수 있다. 따라서 한국어에서 강세와 밀접한 관계가 있는 계수는 F_0이며, 그 다음으로 지속시간, 크기로 나타나 있다. 강세 위치에서 보면 한국어는 음절단위로 부여되고, 강세위치는 구 또는 문장의 전반부와 중반부에 있는 두 음절로 구성된 단어에서는 두 번째 음절에 약간의 강세가 있으며, 문장의 종반부에 있는 단어는 첫 음절에 강세가 있다 그리고 세음절로 구성된 단어에서는 두 번째 음절에 강세가 위치한다.

Ⅲ-2. 억양

문장단위의 음성에서 피치가 움직이는 방향과 강세를 받는 음절간의 상대적인 음높이와 같은 피치의 변화을 '억양'이라 정의하는데, 억양은 피치의 변화로만 구성되는 것이 아니라 지속시간, 강세, 리듬, 속도, 목소리의 음질 등의 음향적 요소가 결합하여 만들어진 것이라고 할 수 있다. 그러나, 일반적으로 억양은 주로 피치로 실현되고 경우에 따라 크기나 지속시간이 수반된다. 한국어에 있어서 문장의 억양은 의미를 변화시키지는 않지만 문장의 형태, 구문구조, 화제의 변경, 의미, 감정 등에 따라 다양한 형태로 나타난다. 문장 중간에서 실현되는 억양은 억양구내에서 그 구의 구문론적, 의미론적인 정보를 나타내어 준다. 예로 한국어의 문장끝의 억양구 단위에서 피치 곡선은 평서문, 의문문, 감탄문은 오르거나 내리거나 평탄하거나 하는 한 가지 방향만의 피치를 가지고 있다[6][10].

Ⅲ-3. 운율구

화자의 자연스러운 발성에 따라 형성되는 단위로서 피치, 지속시간, 크기, 억양 등의 운율변화로 나타나므로, 문장을 적절한 운율구로 나누어서 분석하므로써 음성합성의 자연성을 크게 향상시킬 수 있다. 문장내에서 운율구의 경계를 추출하여 구 단위로 분석하여 문장단위의 여러 가지의 운율 현상을 세분하여 합성음의 자연성을 증가시킬 수 있다.

운율구는 문장내에서 화자와 청취자가 모두 객관적으

로 동의할 수 있는 음성학적 끊김이 있는 단위로서, 일정 수치 이상의 지속시간의 증가나 피치의 변동으로 경계지어진 단위이다. 즉 분명한 끊어 읽기가 이루어진 큰 음성학적 단위의 설정이 바로 운율구라고 하였다. 운율구 한 개에 포함되는 음절수는 그 빈도에 있어 대략 5-8음절 정도가 대종을 이룬다고 할 수 있다[1][3].

Ⅲ-4. 경계의 정의

자연스러운 합성음을 생성해내기 위해서는 먼저 하나의 문장이 들어 왔을 때 이를 어떻게 몇 개의 경계로 나누어 줄 것인가 하는 것이 문제이다. 일반적인 경계는 단어경계와 구,절의 경계로 나눌 수 있지만, 인간이 말을 할 때는 언어의 통사 의미 구조에 관한 지식, 문장의 길이, 말의 속도, 심리적 생리적 요인 등 다양한 요인을 가지고 자연스러운 말의 패턴을 만들어내기 때문에, 문장내에서 구의 경계는 마침표 및 확실한 쉼표구간을 경계로 삼고, 음절수가 많아짐에도 확실한 경계가 없는 경우에는 blank구간을 경계를 나누어서, 이 경계구간으로 연속음의 운율변화곡선을 발생시키는 신경망을 학습시키고자 한다.

Ⅳ. 언어자료 구축 및 음성시료 채집

한국어 음성합성의 방법은, 먼저 음절, 반음절, diphone, 음소 등 음성의 기본단위를 선택하여, 이것을 컴퓨터 내부에 데이터베이스로 구축한후, 이들 중 필요한 것들을 꺼내어 연속시킴으로써 합성음을 만들어 낸다. 여기서 어떠한 합성단위를 사용하는가에 따라, 합성알고리듬의 복잡성과 데이터베이스의 개수와 합성음의 음질에도 큰 영향을 준다.

먼저 합성단위로써 문장, 구, 단어를 사용할 경우, 데이터베이스 내부에 음성이 갖는 거의 모든 정보를 포함하고 있으므로 훌륭한 음질과 매우 자연스러운 합성음을 만들어 낼수 있으나, 많은 양의 데이터베이스가 있어야 함으로, 임의의 한글 텍스트를 음성으로 변화시키는 한국어 음성합성 시스템에서 사용하기 어렵다.

음소를 합성단위로 사용할 경우, 19개 자음과 21개의 모음을 합하여 40개의 데이터베이스만 있으며 음성합성을 할 수 있다. 또한 변이음을 고려하여도 데이터베이스의 수가 많지 않으므로, 적은 메모리로써 하드웨어를 구성할 수 있으며, 데이터베이스를 제어하기가 쉬운 장점이 있다. 그러나 합성단위로서 음소를 이용할 경우, 음소와 음소를 접합시켜야 하는데, 이 과정에서 연결부위의 스펙트럼 포락상에서 불연속점이 발생하여 합성음의 명료도를 손상시킨다[20].

음절은 음소의 결합으로 된 가장 기본적인 음소론적 단위로써, 합성단위로써 음절을 이용할 때, 음질의 면에서도 음절을 구성하고 있는 음소들 사이에 존재하는 변화부분의 정보를 모두 가지고 있으므로, 다른 합성단위과 비교하여 볼 때 좋은 음질의 합성음을 만들 수 있다[19], 그러나 데이터베이스의 수가 많으므로, 변이음 처리, 지속시간조절, 각 데이터베이스와 관련된 계수의 조절 등 데이터베이스를 제어하기 쉽지 않은 단점이 있다.

반음절은 음절 중 2음소형의 데이터베이스만 가지고 있다. 따라서 3음소형의 음절은, CV형 음절과 VC형 음절을 모음 부분에서 끊은 후 성도 필터의 계수를 스므딩과 내삽을 이용하여, 두 음절을 연결하여 만들어낸다. 따라서 이로 인하여 음절 데이터베이스보다 합성음의 음질이 다소 떨어진다. 데이터베이스의 수가 음절 데이터베이스보다 작은 메모리로써 음성합성 시스템을 구성할 수 있으며, 음절에 비하여 데이터베이스를 제어하기 쉬운 장점을 있다.

신경망을 훈련시키기 위한 음성시료로는 고립단어군에서 강세등과, 음소균형 문장군에서의 억양등 그리고 삽입음절 연결어에 의한 변화등을 고려하여, 특정화자로 하여금 언어자료를 발성하게 한 음성시료의 운율을 분석하여 신경망을 훈련시켜 자연음의 운율과 유사한 합성운율을 발생시키고자 한다.

Ⅴ. 신경망 구성 및 훈련

신경망 시스템은 계층적 구조인 입력층, 은닉층, 출력층의 3개의 층으로 구성하였다. 입력층은 7개의 음소열을 나타내는 유니트들로 이루어져 있는데 각 음소는 16개의 유니트로 구성하고, 각종 규칙을 적용한 언어자료를 가지고, 초성 자음 18개, 중성 모음 21개, 종성 자음 7개 및 마침표, 쉼표, 그리고 blank를 음소로 하여 신경망을 훈련시켜 자연음의 운율과 유사한 합성운율을 발생시킨다.

Ⅴ-1 입력패턴

먼저 분절요인에 의한 영향을 고려해주기 위해 전후 3음소를 동시에 입력시키고, 전처리로써 운율구에 대한 각 단어를 초, 중, 종성 및 경계구간을 분리하여 문-음소변환 알고리듬을 사용하여 음운 변동을 적용한 후 이를 통해 얻어진 음소, 즉 자음 18개, 중성 모음 21개, 종성 자음 7개 및 경계구간을 나타내는 마침표와 쉼표, 그리고 blank(□)도 하나의 음소로 하여 6 bit로 할당하고, 운율구내에서의 음소의 상대적인 위치에 관한 정보와

총음소에 대한 정보를 각각 5 bit로 할당하여, 총16비트의 신경망의 입력 패턴으로 구성하였다.

다음은 아래 문장에서 예로 /노래를/에 대한 각 음소 /ㄴ/ /ㅗ/, /ㄹ/, /ㅐ/, /ㄹ/, /ㅡ/ /ㄹ/에 대해 1진 벡터로 표현한 예이다.

"민수는 노래를 하고, 영이는 음악을 듣는다."
문장내의 /노래를/에 대한 입력패턴 예

ㄴ[000010 01010 10110]
ㅗ[010100 01011 10110]
ㄹ[000100 01100 10110]
ㅐ[011001 01101 10110]
ㄹ[000100 01110 10110]
ㅡ[010111 01111 10110]
ㄹ[000100 10000 10110]

이렇게 음소 기호열로 구성된 입력패턴은 총 112개의 비트열로 구성되는데, 입력층에 매 패턴제시마다 전후 3음소씩 모두 7음소가 학습에 참여하게 된다. 학습에 참여하는 7개의 음소 중 실제로 학습이 이루어지는 음소는 네 번째 음소이며, 각각의 음소들은 순차적으로 쉬프트 되면서 학습이 이루어지게 된다.

V-2. 출력패턴

각 음소의 지속 시간 동안에 피치와 에너지 변화에 대한 출력 패턴은 회귀 분석을 통한 추세선을 이용하여 3차 다항식으로 근사화한 후 그 계수인 p_3, p_2, p_1(e_3, e_2, e_1)과 초기 피치값 p_0(e_0)를 1진 벡터화 하여 작성하였다. 그리고 다항식의 변수인 지속 시간 d은 그 프레임수를 2진부호화 하여 출력 패턴을 작성하고, 무성자음의 경우 피치가 존재하지 않으므로 피치변화곡선의 계수와 초기 피치값은 모두 0으로 부호화하였으며 지속시간은 모음과 유성자음의 경우와 마찬가지로 그 프레임수를 1진 벡터화하여 작성한다.

이때 각 음소의 지속시간과 초기 피치 및 에너지값의 범위는 각각 0~300 msec (0~24 frame), 0~6.0로 하고 이를 1진 벡터화하기 위해서 각각 41bit씩을 할당하였다. 운율 변화 곡선의 다항식 계수들은 그 최대 존재 범위가 0~9이므로 이를 부호화하기 위해서 지속시간, 초기 피치 및 에너지값과 마찬가지로 41 bit를 할당하여 첫 1 bit는 부호, 다음 10 bit는 단 자리를 다음 10 bit는 소숫점 첫째 자리를 다음 10 bit 는 소숫점 둘째 자리를 그리고, 다음 10 bit는 소숫점 셋째 자리를 나타내도록 한다[4].

VI. 결론

자연성이 부가된 문장단위의 합성음에 대한 음성데이타의 효율적 분석을 위해서는 각 문장의 내부를 음성적으로 의미있는 단위로 끊어주고, 개별적 언어자료구축에 따른 중복투자를 줄이고, 각종 알고리즘을 적절히 비교 평가하기 위해서 데이터량을 줄이면서 실제 한국어의 발성에 나타나는 음운현상을 가능한 많이 포함하며, 특정 태스크에 집중되지 않는 것이 바람직하다. 이러한 언어자료를 가지고 단어의 음소성분과 문장의 마침표, 쉼표 그리고 blank의 변화성분과 마찬가지로 음소로 취급하여 신경망을 학습시키고, 해당음소의 운율구내에서의 상대위치에 따른 운율 변화도 같이 훈련시켜, 복잡한 운율법칙 알고리듬을 위한 프로그램 작성없이 연속음의 운율변화곡선을 발생시키는데 필요한 신경망을 훈련시키는 방법을 제안하였다.

[참 고 문 헌]

[1] Eric Sanders and Paul Taylor, "Using Statistical Models to Predict Phrase Boundaries for Speech Synthesis." in EU ROSPEECH'95 Spain, 1995.

[2] 임 운천, "한국어 법칙합성을 위한 운율법칙 구현에 관한 연구", 서울대학교 박사학위논문, 1991.

[3] 성철재, "한국어 리듬의 실험음성학적 연구", 서울대학교 박사논문, 1995.

[4] 류창수, "신경망 합성에 따른 운율 제어기 성능 비교에 관한 연구", 호서대학교 석사논문, 1998

[5] 김현준, "신경망을 사용한 문-변이음 변환에 관한 연구", 호서대학교 석사논문, 1998.

[6] 김선철, "국어 억양의 음성학·음운론적 연구", 서울대학교 박사논문 1996.

[7] 김연준, 오영환 "한국어 문서-음성 변환 시스템에서의 구문분석에 의한 운율조절에 관한 연구" 제 10회 음성통신 및 신호처리 워크샵 논문집, 1993.

[8] 허 준, "무제한 단어 한국어 음성합성 시스템에서의 운율정보 구현에 관한 연구", 서울대학교 석사학위 논문, 1990.

[9] Ostendorf, "Parse scoring with prosodic information : an analysis and synthesis approach." in Computer Speech and Language. July 1993.

[10] 이현복, "음성학과 언어학", 서울대학교출판부,1996

[11] 정국외 4, "음성인식/합성을 위한 국어의 음성-음운론적 특성 연구" 한국 음향학회지 제 13권 6호,1994.

[12] J. Allen, M. S. Hunnicutt & D. Klatt , From Text To Speech : The MITalk System.Cambridge University Press, 1987.
[13] D. H. Klatt, "Structure of phonological rule component for synthesis by rule program", IEEE Vol.ASSP-24 No.5, pp.391-398, 1976.
[14] D. O'Shaughnessy, "Automatic Speech Synthesis", IEEE Communication magazine, pp26-34, 1983. 12.
[15] Do-Heung Ko, Declarative intonation in Korean ; An acoustical study of F0 declination, Ph. D dissertation, Univ. of Kansas, 1988.
[16] N. Umeda, "Linguistic rules for text-to-speech synthesis", Proc. of IEEE, vol. 64, No. 4, pp. 433-451, Apr. 1976.
[17] R. P. Lippmann, "An Introduction to Computing with Neural Nets", IEEE ASSP Magagine, Vol. 4, No. 2, pp. 4-22, April 1987.
[18] J. M. Zurada, Introduction to Artificial Neural Systems, West Publishing Company, 1992.
[19] 허웅, 국어 운운학, 정음사, 1985
[20] H.Dettwealer and W.Hess, "Concatenation rules for demisyllable speech synthesis," NATO ASI Series, vol.F16, 1985